Mining and Classification of Neologisms in Persian Blogs

Karine Megerdoomian
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102, USA
karine@mitre.org

Ali Hadjarian
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102, USA
ahadjarian@mitre.org

Abstract

The exponential growth of the Persian blogosphere and the increased number of neologisms create a major challenge in NLP applications of Persian blogs. This paper describes a method for extracting and classifying newly constructed words and borrowings from Persian blog posts. The analysis of the occurrence of neologisms across five distinct topic categories points to a correspondence between the topic domain and the type of neologism that is most commonly encountered. The results suggest that different approaches should be implemented for the automatic detection and processing of neologisms depending on the domain of application.

1 Introduction*

Since its beginnings in 2001, the Persian blogosphere has undergone a dramatic growth making Persian one of the top ten languages of the global blog community in 2007 [Sifry 2007].

One of the main challenges in the automatic analysis and processing of Persian language blogs is the accelerated emergence of neologisms in online discourse. These newly created words that cannot be found in traditional lexicons primarily consist of adopted English loanwords, such as دانلود *dânlod* 'download' or آنلاین *ânlâyn* 'online', and innovative constructions based on Persian word-formation principles, as in فیلترشکن *filtershekan* 'anti-filter software' or چتیدن *chatidan* 'to chat'.

In this paper, we investigate the distinct classes of neologisms encountered in Persian language blogs. Since the main goal of the project is to build a topic classification system for blogs, we focused on extracting neologisms that would have the most discriminatory power in distinguishing between the various classes.

For the purposes of this study, we collected a corpus of Persian language blogs from five different topic categories of sports, medicine, politics, Internet, and cinema. The neologisms are automatically extracted by running the documents through a morphological analyzer. Since these new word coinages are not recognized by the analyzer, they are tagged as unknowns. A weight-ordered list of unknown words is then generated for each topic category, using information gain as the measure, as described in Section 2.3. The more significant neologisms for each category are then manually identified from the generated weight-ordered lists, focusing on the top 200 words, and classified based on their linguistic characteristics. The results indicate that the type of neologism found in the blog posts in fact corresponds to the topic domain. Hence, for example, while English loans are highly prominent in technical and Internet related posts, new morphological constructions are more common in the domain of politics. Building on these results, we argue that distinct approaches are required for processing the adopted loan words and the morphologically constructed neologisms.

* This research is part of a larger project on the study of Persian language blogs supported by a Mission-Oriented Investigation and Experimentation (MOIE) program at MITRE.

Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity, pages 6–13,
Los Angeles, California, June 2010.

2 Extraction Process

2.1 Blog Data

The blog data for this study comes from Blogfa[1], a popular Persian blog hosting site. The topic index provided by Blogfa itself has allowed us to rapidly collect large amounts of blog data coming from topic categories of interest, by eliminating the need to manually label the data. Table 1 provides a list of the five topic categories used in this study, as well as the total number and the median size of the collected blogs for each. The table also includes the average number of words in each topic category. The blogs were collected in February 2010, starting with the most recent blog posted in each topic and moving back chronologically.

topic category	# of blogs	median size	average # of words
Internet	497	14 kb	986
Cinema and theatre (*sinama va ta'atr*)	255	18 kb	1380
Political (*siyasat-e-rooz*)	500	22 kb	2171
Medical (*pezeshki*)	499	27 kb	2285
Sports (*varzesh*)	498	19 kb	1528

Table 1 – Topic categories of interest and the total number, median size, and average length of the collected blogs for each topic

2.2 Linguistic Parsing

The collected documents are run through a Persian morphological parser that analyzes all word forms including compounds and provides a part of speech tag for the valid analyses [Amtrup 2003]. The morphological analyzer was developed for use in a Persian-English machine translation system and provides part of speech as well as all syntactically relevant inflectional features for a word [cf. Megerdoomian 2000]. The morphological formalism consists of a declarative description of rules utilizing typed feature structures with unification. The morphological analysis component takes advantage of a lexicon of about 40,000 entries in citation form that had been developed in the period of 1999-2002 for coverage of online news articles and includes nouns, adjectives, verbs, adverbs and closed class items. In addition, there are about 5,000 common proper noun entities listed in the lexicon. After morphological analysis, dictionary lookup eliminates all erroneous analyses. Any element that is not successfully associated with a word in the lexicon is tagged as an unknown.

The current morphological analyzer has a coverage of 97% and an accuracy of 93% on a 7MB corpus collected from online news sources. The system fails to analyze conversational forms. Other unanalyzed tokens are mainly proper nouns and words missing in the lexicon.

2.3 Information Gain

To automatically detect the most pertinent unknown terms per blog category, we employ an *information gain* (IG) based feature selection approach. IG's effectiveness as a feature selection measure for text topic categorization, the ultimate objective of this project, has been well studied [Yang and Pedersen 1997].

Information gain is a statistical measure for calculating the expected reduction in *entropy* [Mitchell 1997]. Entropy, a common measure in information theory, captures the impurity of a collection of examples relative to the intended classification. For a binary classification task, the entropy of a set of examples E is defined as:

$$Entropy(E) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

where p_+ is the proportion of positive examples and p_- is the proportion of negative examples in E. Moreover, it is assumed that $0 \log_2 0$ is equal to zero. The discriminatory power of an attribute for a given class can then be measured by IG, which is the reduction in entropy caused by the partitioning of the data using that attribute. The IG for example set E and attribute A is defined as:

$$IG(E, A) = Entropy(E) - \sum_{v \in Values(A)} \frac{|E_v|}{|E|} Entropy(E_v)$$

where $Values(A)$ represents the set of all possible values that attribute A can take on and E_v is the set of examples for which attribute A takes on value v. In this study, each attribute has a binary value which signifies the presence or absence of a given unknown term in the document. So for the purpos-

[1] www.blogfa.com

es of this study, the above equation can be formulated as:

$$IG(D,t) = Entropy(D) - \frac{|D_t|}{|D|} Entropy(D_t) - \frac{|\overline{D_t}|}{|D|} Entropy(\overline{D_t})$$

where D is the set of all documents, t is a given term, D_t is the set of documents in which term t occurs, and $\overline{D_t}$ is the set of documents in which term t does not occur.

Translit.	Weight	Translation
vyndvz	0.100033	Windows
danlvd	0.080559	download
fayl	0.058319	file
karbran	0.051595	users
Java	0.048287	Java
klyk	0.048180	click
yahv	0.044999	Yahoo
nvkya	0.044807	Nokia
flG	0.042718	Flash
mrvrgr	0.041374	browser
hk	0.041074	hack
msnJr	0.040853	Messenger
Ct	0.039987	chat
psvrd	0.039213	password
kd	0.035936	code

Table 2 –The top weighted unknown terms for the Internet topic category and their associated information gain

Since the aim of feature selection for this paper is that of identifying the most pertinent unknown terms for each topic category, an additional constraint is imposed on the feature selection process. Here, for a term to be selected, it not only needs to have a high IG, but it needs to be present in a higher proportion of positive examples than the negative ones. This prevents the selection of terms that while are good descriptors of the negative class and thus carry a high IG, are not necessarily pertinent to the positive class (i.e., the topic category under consideration). So IG of a term not meeting the above constraint is effectively set to zero.

As indicated previously, the 200 unknown terms with the highest IG for each topic category are thus selected for the analysis portion of this study. Table 2 depicts a sample set of the top weighted terms for the Internet category in transliteration and translation. The transliteration schema was designed to display the results of the morphological analyzer system. It directly represents the Persian script and provides a bijective, one-to-one mapping of the characters. The transliteration omits any diacritics, including vowels, that are not represented in the original script.

2.4 Candidate List

The weight-ordered list of unknown words provides a candidate list of potential neologisms. However, the set of unknown terms extracted from each topic category includes proper names, spelling errors, conversational language forms and neologisms. We therefore manually study the candidate list in order to identify the appropriate classes of neologisms. The results are classified based on the observed linguistic characteristics and a quantitative analysis is performed for each topic category.

3 Neologism Classification

Persian language blogs include a large number of neologisms ranging from new usages in conversational language to newly coined words to designate new technology or political concepts. We performed a qualitative, linguistic investigation of Persian language blogs, consisting of posts from four main categories of technical, political, arts, and personal diary [Megerdoomian 2008]. The goal of this study was to identify elements of Persian Blogspeak that indicate language change and which fail to be analyzed by the existing Persian machine translation and information extraction systems that had been developed for online news sources. The study pointed to four main categories of new word formation found in Persian language blogs:

- Borrowings (mainly from English and French)
- Compounding
- Affixation
- Conversion: Semantic and functional shift

These neologisms were identified based on the prevalent linguistic classification of newly formed words (see for instance the classification of neologisms described in [Grzega and Schoener 2007]).

These four classes of neologisms are described in more detail in the rest of this section.

3.1 Borrowings

A large number of new loan words can be found in blogs. Although they may sometimes be inserted within the text in the original language, they are generally transcribed into Persian script. These loans are used as regular words and can take Persian inflectional affixes. Some examples are provided in Table 3.

Persian	Transcription	Translation
مونیتور	*monitor*	Monitor
فیلترینگشون	*filteringeshun*	their filtering
سایتها	*sâythâ*	sites
پسوردتان	*pasvordetân*	your password
وایرلس	*vâyerles*	Wireless
تایم‌لاین	*tâym lâyn*	Timeline
سکسوالیته	*seksuâliteh*	Sexuality
نوستالژی	*nostâlji*	Nostalgy

Table 3 – Loan words in Persian blogs

An analysis of the occurrence of loans with respect to the various topic domains shows that the Internet category contains a large number of English language loans, whereas the more established scientific domain of medicine tends to use French borrowings. Also within the same category, new technological or scientific additions are generally expressed in English. For instance, in the cinema category, most technical words are of French origin – e.g., اکران from *écran* 'screen' or تیتراژ from *titrage* 'opening credits'. However, new loans have entered the field from English, e.g., انیمیشن *animeyshen* 'animation'.

3.2 Compounding

Compounding is a productive word-formation process in Persian and refers to cases where two or more roots are combined to form a new word. Examples of compounding include راهکار *râhkâr* (consisting of *râh* 'path' and *kâr* 'work' and now being used to mean 'guideline' or 'solution'); سربرگ *sarbarg* (from *sar* 'head' and *barg* 'leaf, piece of paper' signifying 'letterhead'); and دگرباش *degarbâsh* (formed with *degar* 'other' and *bâsh* 'being', meaning 'queer'). In many cases, however, one of the roots is a borrowing that is combined with a Persian root form. Examples of this multilingual compounding construction include تابوسازی *tâbusâzi* (taboo + to make) 'making taboo' and لینکدونی *linkduni* (link + storage) meaning 'blogroll'.

Recently, there has been a concerted effort by the Persian Language Academy to replace borrowings from foreign languages by equivalent Persian constructions. Thus, the traditional هلیکوپتر *helikopter* 'helicopter' has been replaced by بالگرد *bâlgard* by combining Persian *bâl* 'wing' and *gard* 'turn'. Similarly, the French loanword سناریو *senâryo* 'screenplay' is now being replaced by فیلمنامه *filmnâmé* composed of *film* and *nâmé* 'letter, book'.

Persian has a very productive compounding strategy for forming new verbs where a nominal or adjectival root is combined with a light verb, i.e., a verb that has lost some of its semantic value. Many new verbs, especially in the technical domain, are formed following this construction as illustrated in Table 4.

Persian	Transcription	Translation
کلیک کردن	*kelik kardan*	to click
چت کردن	*chat kardan*	to chat
اس‌ام‌اس زدن	*es-em-es zadan*	to send a text message
کنسل شدن	*kansel shodan*	to be cancelled

Table 4 – Compound verb formation

3.3 Affixation

New words are often created following a productive word-formation pattern using suffixes. For instance, the agentive suffix *–gar* is used to form مرورگر *morurgar* 'browser' by combining with *morur* 'review', and فتنه‌گر *fetne-gar* 'seditious' when combined with *fetne* 'sedition'[2]. Another common affix used to form new words is *–estân* which indicates a place. This suffix can be found in terms like وبلاگستان *veblâgestân* (weblog + *-stan*) 'blogosphere' or لینکستان *linkestân* (link + *-stan*) 'blogroll'.

In addition to the compound verb formation, bloggers have started creating simple verbs by combining the verbal ending *–idan* with nominal

[2] *Fetne-gar* is a relatively new creation that is used alongside the traditional *fetne-ju* 'seditious'. There is a clear sense among native speakers, however, that *fetne-gar* refers to a person who is more agentive, actively causing discord.

roots as in چتیدن *chatidan* ‘to chat’ or لاگیدن *lâgidan* ‘to blog’.

3.4 Conversion

Another type of neologism found in Persian language blogs consists of existing words that are being used in a new context, bringing about a semantic shift. In certain instances, the part-of-speech category may also shift. One example is the adjective شفاف *shafâf* ‘transparent’ that is being used more and more frequently as an adverb in political contexts with the meaning ‘openly, transparently’.

This category, however, is difficult to detect automatically with the methodology used since these words already exist in traditional lexicons and are not tagged as unknowns by the morphological parser. Identifying conversions and semantic shifts currently requires a manual exploration of the data; we leave a study of this class of neologisms for future work.

4 Topic and Neologism Correspondence

An investigation of the neologisms for each topic category clearly suggests that there is a close relationship between the class of neologisms and the topic domain.

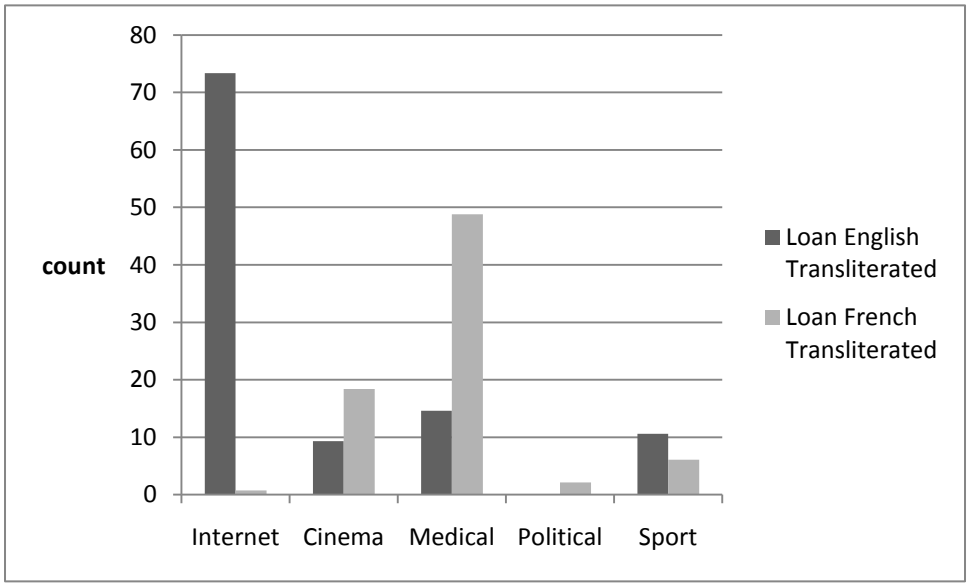

Figure 1 – Loan neologisms across topic categories

Starting from the weight-ordered candidate list for each topic category, we manually examined and labeled each unknown word according to the neologism classification described in Section 3. In order to identify the correct class, each unknown word was considered within its context of occurrence in the corresponding blog posts and classified according to the usage within the sentence. In addition, proper names, conversational forms of existing words, and spelling errors were tagged separately.

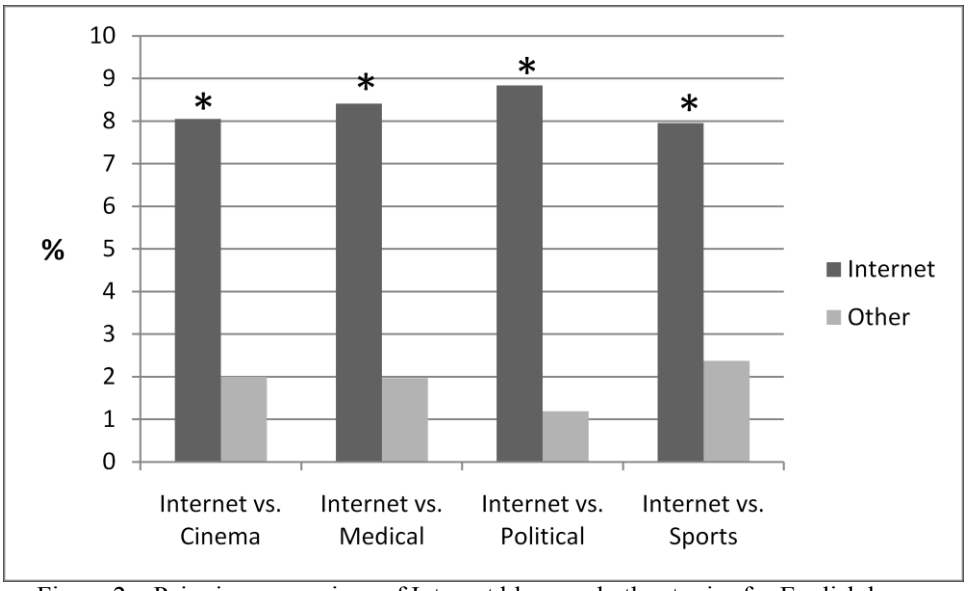

Figure 2 – Pairwise comparison of Internet blogs and other topics for English loans

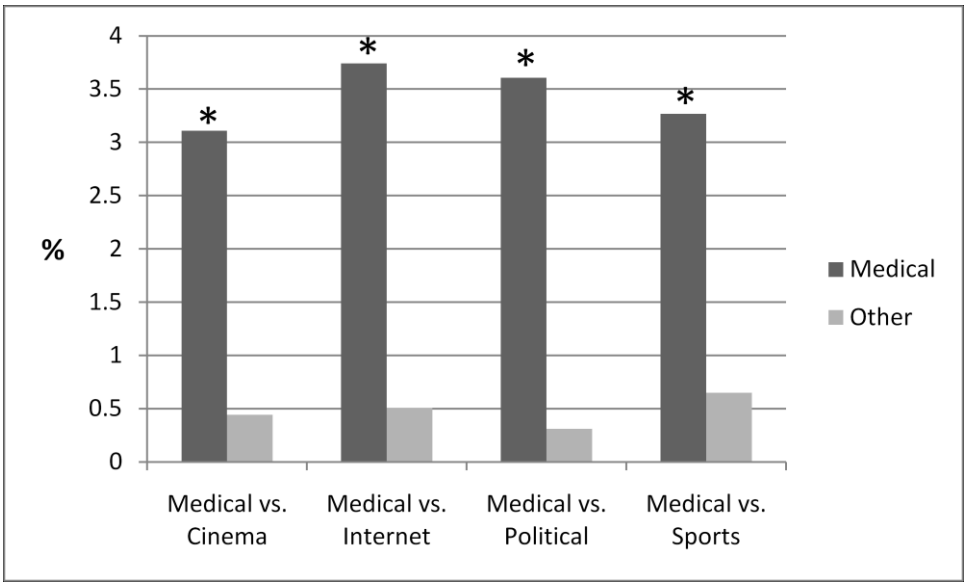

Figure 3 – Pairwise comparison for Medical blogs and other topics for French loans

Figure 1 illustrates the correspondence of the number of borrowings per topic category in the corresponding candidate list. The results show that the most common way of forming new words within blogs dealing with Internet and computer technology is to use borrowings from English. In the medical domain, however, most neologisms are scientific terms borrowed mainly from the French language. The domain of cinema and theatre also includes a large number of loans from French. However, most of the French loans across topics seem to be older borrowings while the newer loans (i.e, within the last three to five years) are almost always from the English language. A statistical analysis of the results indicate that these correspondences are significant as shown in Figure 2 for English loans and in Figure 3 for French loans. Figure 2 illustrates a pairwise comparison between the Internet category and other blog topics based on the average percentage of documents in which a given term from the English loan neologism category is included. (*) indicates a statistically significant difference between the two percentages. Figure 3 shows a similar result for the pairwise comparison between the Medical category and other topics for the French loan class.

Figure 4 shows the relative usage of affixation and compounding strategies for the creation of new words. Although affixation is used to some degree in both the Internet and medical domains, they do not occur as often as the English or French loans (cf. Figure 1 above). Interestingly, the blogs that fall within the political topic category do not make much use of borrowings from English and French. Instead, they tend to create new words by applying productive affixation and compounding strategies. In most instances, the words used to form neologisms in the politics category are based on words of Arabic and Persian origin. Figure 5 illustrates the pairwise comparison between the Political and other blog topics based on the average percentage of documents in which a given term from the affixation and compounding class of neologisms is included. (*) indicates a statistically significant difference between the two percentages.

Hence, while the Internet blogs make heavy use of English loans in the creation of new words, political blogs tend to use affixation and compound strategies for word-formation. These results sug-

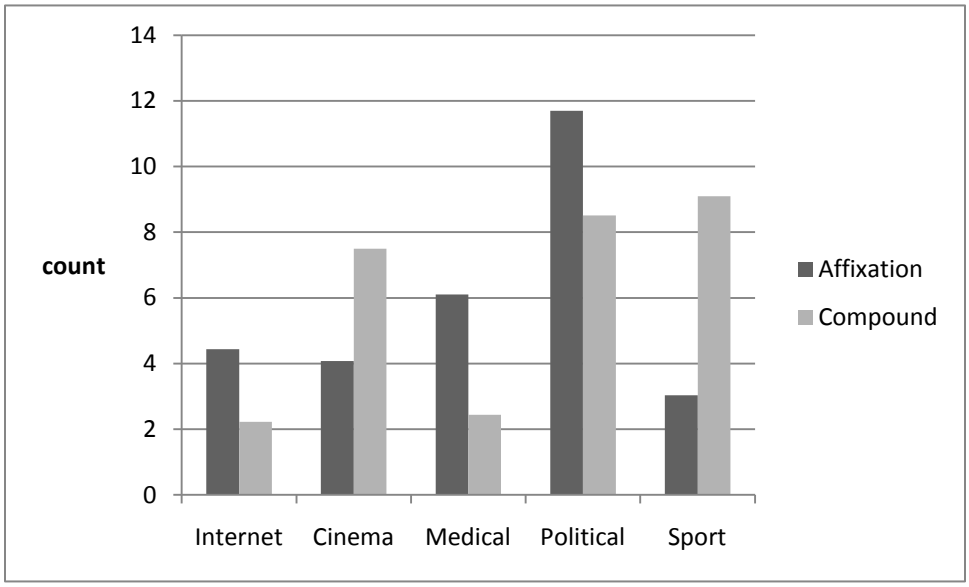

Figure 4 – Affixation and compounding strategies for the creation of new words across various blog topics

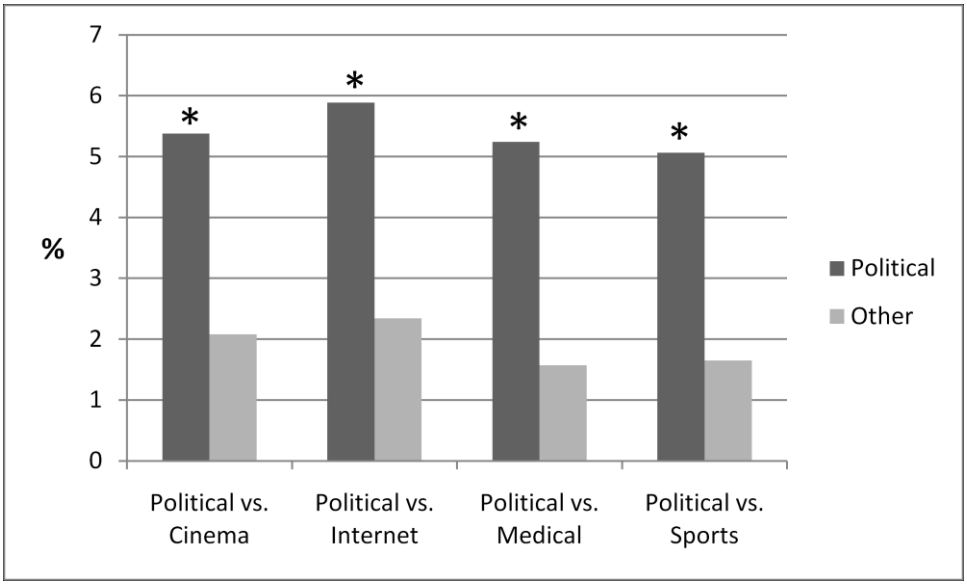

Figure 5 – Pairwise comparison for Political blogs and other topics for affixation and compounding

gest that, depending on the domain of interest for the particular NLP application, distinct methodologies for the automatic detection and processing of neologisms should be implemented.

5 Conclusion

This paper presents an investigation of neologisms in Persian blog posts across five distinct topic areas. We employ morphological analysis in conjunction with a profile-based classification technique to successfully extract a pertinent candidate list for identifying new word-level constructions in blogs. These neologisms are then classified based on their linguistic characteristics and word-formation strategies and the quantitative analysis points to a significant correspondence between neologism classes and blog topic domains.

Based on these results, we propose that the detection and processing strategies should be tailored to the domain of the NLP application for greater efficiency. In particular, a derivational morphological system can be developed by implementing the productive affixation and compounding rules used in Persian word formation. This system can be used to extend the existing analysis and translation systems in the domain of politics. Loans from English, on the other hand, can be automatically processed by using previously implemented methodologies for transcribing Persian script into the English writing system [Megerdoomian 2006, Johanson 2007]. Such a system would be beneficial in recognizing the large number of loans encountered in the technical and scientific domains.

This work is part of a larger project for automatic topic classification and sentiment analysis in Persian language blogs. We extract the most pertinent neologisms encountered in the blog corpus in order to enhance the topic classification system. In addition, the results obtained will be used to extend the current morphological parser to improve coverage and identification of newly formed words.

References

Amtrup, Jan W. 2003. Morphology in machine translation systems: Efficient integration of finite state transducers and feature structure descriptions. *Machine Translation,* 18(3), pp. 217-238.

Grzega, Joachim and Marion Schoener. 2007. *English and general historical lexicology: Materials for onomasiology seminars.* Onomasiology Online Monographs, Vol. 1. Germany.

Ittner, D.J., Lewis, D.D., and Ahn, D.D. (1995). Text categorization of low quality images. In Symposium on Document Analysis and Information Retrieval. Las Vegas, NV.

Johanson, Joshua. 2007. Transcription of names written in Farsi into English. In *Proceedings of the Computational Approaches to Arabic Script-based Languages* (CAASL2). LSA Linguistic Institute, Stanford.

Kelly, John and Bruce Etling (2008). Mapping Iran's online public: Politics and culture in the Persian blogosphere. Research Publication No. 2008-01, The Berkman Center for Internet and Society at Harvard Law School. April 6.

Megerdoomian, Karine. 2008. Analysis of Farsi weblogs. MITRE Tech Report 080206. August 2008.

Megerdoomian, Karine. 2006. Transcription of Persian proper name entities into English. Technical report, Inxight Software, Inc.

Megerdoomian, Karine. 2000. Unification-based Persian morphology. In *Proceedings of CICLing 2000*. Alexander Gelbukh, ed. Centro de Investigacion en Computacion-IPN, Mexico.

Mitchell, Tom M. 1997. *Machine learning*. McGraw-Hill.

Pacea, Otilia. 2009. New worlds, new words: On language change and word formation in Internet English and Romanian. In *The annals of Ovidius University Constanta- Philology*, issue 20, pp: 87-102.

Salton, G. 1991. Developments in automatic text retrieval. Science, v.253: 974-980.

Sebastiani, F. 2002. Machine learning in automated text categorization. In ACM Computing Surveys, 34(1): 1-47

Sifry, Dave. 2007. The Technorati state of the live web: April 2007.

Yang, Yiming and Jan Pedersen, J.O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of International Conference on Machine Learning.*